Supplementary Material to "Langevin Incremental Mixture Importance Sampling"

Matteo Fasiolo, Flávio Eler de Melo and Simon Maskell

April 7, 2017

1 Derivation of differential equations (5) and (6)

Firstly, define $\boldsymbol{\mu}_t = \mathbb{E}(\mathbf{x}_t)$ and $\boldsymbol{\Sigma}_t = \mathrm{Cov}(\mathbf{x}_t)$. Then consider the discrete time iteration for the mean

$$\boldsymbol{\mu}_{t+\delta t} = \boldsymbol{\mu}_t + \frac{\delta t}{2}\nabla \log \pi\{\boldsymbol{\mu}_t\},$$

then

$$\frac{1}{\delta t}(\boldsymbol{\mu}_{t+\delta t} - \boldsymbol{\mu}_t) = \frac{1}{2}\nabla \log \pi(\boldsymbol{\mu}_t).$$

letting $\delta t \to 0$ leads to (5). Now consider the discrete time iteration for the covariance matrix

$$\boldsymbol{\Sigma}_{t+\delta t} = \left[\mathbf{I} + \frac{\delta t}{2}\nabla^2 \log \pi(\boldsymbol{\mu}_t)\right]\boldsymbol{\Sigma}_t\left[\mathbf{I} + \frac{\delta t}{2}\nabla^2 \log \pi(\boldsymbol{\mu}_t)\right]^T + \delta t\mathbf{I},$$

by expanding the r.h.s., and noticing that $\nabla^2 \log \pi(\boldsymbol{\mu}_t)$ is symmetric, we obtain

$$\frac{1}{\delta t}(\boldsymbol{\Sigma}_{t+\delta t} - \boldsymbol{\Sigma}_t) = \left\{\frac{1}{2}\nabla^2 \log \pi(\boldsymbol{\mu}_t)\right\}\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}_t\left\{\frac{1}{2}\nabla^2 \log \pi(\boldsymbol{\mu}_t)\right\} + \mathbf{I} + \frac{\delta t}{4}\nabla^2 \log \pi(\boldsymbol{\mu}_t)\nabla^2 \log \pi(\boldsymbol{\mu}_t),$$

letting $\delta t \to 0$ leads to (6).

2 Derivation of the Population Effective Sample Size

Consider a Gaussian importance density, with mean $\boldsymbol{\mu}_1$ and covariance $\boldsymbol{\Sigma}_1$, and a Gaussian target, with mean $\boldsymbol{\mu}_2$ and covariance $\boldsymbol{\Sigma}_2$. Then the PESS is defined by

$$\mathrm{PESS}\{\phi(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \phi(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\} = \left\{\int \left[\frac{\phi(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)}{\phi(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}\right]^2 \phi(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\, d\mathbf{x}\right\}^{-1} = \mathbb{E}(w^2)^{-1}.$$

Simple manipulations lead to

$$\begin{aligned}\mathbb{E}(w^2) &= (2\pi)^{-\frac{d}{2}}\frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\boldsymbol{\Sigma}_2|}\int \exp\left\{-(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}d\mathbf{x} \\ &= \left(2^d|-\mathbf{A}|\right)^{-\frac{1}{2}}\frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\boldsymbol{\Sigma}_2|}e^{\mathbf{c}-\frac{1}{4}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}},\end{aligned}$$

where

$$\mathbf{A} = \frac{1}{2}\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}, \quad \mathbf{b} = 2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \quad \text{and} \quad \mathbf{c} = \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2.$$

The exponent can be simplified, in fact the properties $\mathrm{Tr}(\mathbf{X}+\mathbf{Y}) = \mathrm{Tr}(\mathbf{X}) + \mathrm{Tr}(\mathbf{Y})$ and $\mathrm{Tr}(\mathbf{XYZ}) = \mathrm{Tr}(\mathbf{ZXY}) = \mathrm{Tr}(\mathbf{YZX})$ lead to

$$\begin{aligned}
\mathbf{c} - \frac{1}{4}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} &= \mathrm{Tr}\Big(\mathbf{c} - \frac{1}{4}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}\Big) \\
&= \mathrm{Tr}\Bigg\{\Big(\frac{1}{2}\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\Big)^{-1}\Big(- \frac{1}{2}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_2\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1} - \frac{1}{2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1} \\
&+ \frac{1}{2}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\Big)\Bigg\}.
\end{aligned}$$

Then we can use the property

$$\big(\mathbf{X}^{-1} + \mathbf{Y}^{-1}\big)^{-1} = \mathbf{X}\big(\mathbf{X}+\mathbf{Y}\big)^{-1}\mathbf{Y} = \mathbf{Y}\big(\mathbf{X}+\mathbf{Y}\big)^{-1}\mathbf{X}, \tag{1}$$

to obtain

$$\begin{aligned}
\mathbf{c} - \frac{1}{4}\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} &= \mathrm{Tr}\Bigg\{\Big(2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\Big)^{-1}\Big(\boldsymbol{\mu}_1\boldsymbol{\mu}_1^T + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_1\boldsymbol{\mu}_2^T - \boldsymbol{\mu}_2\boldsymbol{\mu}_1^T\Big)\Bigg\} \\
&= \Big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\Big)^T\Big(2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2\Big)^{-1}\Big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\Big),
\end{aligned}$$

after some rearrangements. This leads to

$$\mathbb{E}(w^2) = \Big(2^d|\boldsymbol{\Sigma}_2^{-1} - \frac{1}{2}\boldsymbol{\Sigma}_1^{-1}|\Big)^{-\frac{1}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{-1}\exp\Bigg\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Bigg\}.$$

We can avoid computing the inverses of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ by using (1) to obtain

$$|\boldsymbol{\Sigma}_2^{-1} - \frac{1}{2}\boldsymbol{\Sigma}_1^{-1}|^{-\frac{1}{2}} = \big(2^d|\boldsymbol{\Sigma}_1||2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2|^{-1}|\boldsymbol{\Sigma}_2|\big)^{\frac{1}{2}},$$

so finally

$$\mathbb{E}(w^2) = |\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|^{-\frac{1}{2}}|2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2|^{-\frac{1}{2}}\exp\Bigg\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Bigg\},$$

which exists if $2\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_2$ is positive definite.

3 Mixture of warped Gaussians example details

The gradient and Hessian of the log-density of a general weighted mixture density

$$\pi(\mathbf{x}) = \sum_{i=1}^{r} w_i p_i(\mathbf{x}),$$

are

$$\nabla \log \pi(\mathbf{x}) = \sum_{i=1}^{r} \frac{w_i p_i(\mathbf{x})}{p(\mathbf{x})} \nabla \log p_i(\mathbf{x}),$$

$$\nabla^2 \log p(\mathbf{x}) = \sum_{i=1}^{r} \frac{w_i p_i(\mathbf{x})}{p(\mathbf{x})} \left\{ \nabla^2 \log p_i(\mathbf{x}) + \nabla \log p_i(\mathbf{x}) \nabla \log p_i(\mathbf{x})^T \right\} - \nabla \log p(\mathbf{x}) \nabla \log p(\mathbf{x})^T.$$

The mixture used in the paper is composed of d-dimensional warped Gaussian densities, with parameters a, b, s_1 and s_2. Let $\mathbf{z}$ be a d-dimensional vector such that $z_1 = x_1 - s_1$, $z_2 = x_2 - s_2$ and $z_i = x_i$ for $i = 3, \ldots, d$. Then the entries of $\nabla \log p(\mathbf{x})$ and $\nabla^2 \log p(\mathbf{x})$ can be obtained by noticing that the Jacobian of this transformation is the identity matrix and using

$$\frac{\partial \log p(\mathbf{z})}{\partial z_1} = -\frac{z_1}{a^2} - 2bz_1\{z_2 + b(z_1^2 - a^2)\},$$

$$\frac{\partial \log p(\mathbf{z})}{\partial z_2} = -z_2 - b(z_1^2 - a^2), \quad \frac{\partial \log p(\mathbf{z})}{\partial z_i} = -z_i, \quad \text{for} \quad i = 3, \ldots, d.$$

and

$$\frac{\partial^2 \log p(\mathbf{z})}{\partial z_1^2} = -a^{-2} - 2b\{z_2 + b(z_1^2 - \sigma^2)\} - 4b^2 z_1^2,$$

$$\frac{\partial^2 \log p(\mathbf{z})}{\partial z_1 \partial z_2} = \frac{\partial^2 \log p(\mathbf{z})}{\partial z_2 \partial z_1} = -2bz_1, \quad \frac{\partial^2 \log p(\mathbf{z})}{\partial z_i \partial z_i} = -1, \quad \text{for} \quad i = 2, \ldots, d,$$

with all the remaining entries of the Hessian being equal to zero. In the main text we used six densities with the following parameters

$$\mathbf{a} = \{1, 6, 4, 4, 1, 1\}, \quad \mathbf{b} = \{0.2, -0.03, 0.1, 0.1, 0.1, 0.1\},$$

$$\mathbf{s}_1 = \{0, 0, 7, -7, 7, -7\}, \quad \mathbf{s}_2 = \{0, -5, 7, 7, 7.5, 7.5\},$$

where, for instance, the i-th element of $\mathbf{a}$ is the value of a use to define the i-th warped Gaussian. The weights of the target and of the importance mixture components are

$$\mathbf{w}_{\mathrm{T}} \propto \{1, 4, 2.5, 2.5, 0.5, 0.5\}, \quad \mathbf{w}_{\mathrm{IS}} \propto \{1, 4, 2.5, 2.5\}.$$

4 Bayesian logistic regression details

The gradient the log-posterior is

$$\nabla \log \pi(\boldsymbol{\theta}) = \mathbf{X}^T \mathbf{y} - \sum_{i=1}^{n} \frac{\mathbf{X}_{i:}^T}{1 + e^{-\mathbf{X}_{i:}^T \boldsymbol{\theta}}} - \boldsymbol{\alpha} \odot \boldsymbol{\theta},$$

where $\alpha_1 = 0$, $\alpha_j = \lambda$ for $j = 1, \ldots, d$ and $\odot$ is the Hadamard product. The Hessian is

$$\nabla^2 \log \pi(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \frac{e^{\mathbf{X}_{i:}^T \boldsymbol{\theta}}}{(1 + e^{\mathbf{X}_{i:}^T \boldsymbol{\theta}})^2} \mathbf{X}_{i:} \mathbf{X}_{i:}^T - (\boldsymbol{\alpha} \boldsymbol{\alpha}^T)^{\frac{1}{2}},$$

where $\mathbf{X}_{i:}$ is a column vector including the elements of i-th row of $\mathbf{X}$.

5 Ridge-like model details

The means and standard deviations of the prior are

$$\boldsymbol{\gamma} = \{6, 0.5, 5.5, 0.15, 3.0, 0.6\}, \quad \boldsymbol{\beta} = \{1.3, 0.14, 0.289, 0.029, 0.04, 0.1\},$$

and those of the likelihood are

$$\mathbf{y} = \{7, 0.0525, 2, 4\}, \quad \boldsymbol{\sigma} = \{0.5, 0.00144, 0.01, 0.01\}.$$

Define the vectors $\mathbf{p}'$, where $p'_i = (\gamma_i - \theta_i)/\beta_i^2$ for $i = 1, \ldots, 6$, and $\mathbf{l}'$, where $l'_i = (y_i - \mu_i)/\sigma_i^2$ for $i = 1, \ldots, 4$. The μ_is are defined in the main text. Define also the vector

$$\mathbf{u} = \{1/\theta_5, \theta_4, \theta_6, \theta_2, -\theta_1/\theta_5^2, \theta_3\}.$$

Then the i-th element of gradient of the log-posterior density is

$$\{\nabla \log \pi(\boldsymbol{\theta})\}_i = p'_i + l'_1 \mu_1/\theta_i + l'_{j(i)} u_i,$$

for $i = 1, \ldots, 6$, where $j()$ is a function mapping $\{1, 2, 3, 4, 5, 6\}$ to $\{3, 2, 4, 2, 3, 4\}$.

Now, define the vectors $\mathbf{p}''$ and $\mathbf{l}''$, where $p''_i = -1/\beta_i^2$, for $i = 1, \ldots, 6$, and $l''_i = -1/\sigma_i^2$, for $i = 1, \ldots, 4$. Let $\boldsymbol{\mu}'$ be a 4×6 matrix such that

$$\mu'_{1,i} = \mu_1/\theta_i, \quad \text{for } \; i = 1, \ldots, 6,$$

$$\mu'_{2,2} = \theta_4, \;\; \mu'_{2,4} = \theta_2, \;\; \mu'_{3,1} = 1/\theta_5, \;\; \mu'_{3,5} = -\theta_1/\theta_5^2, \;\; \mu'_{4,3} = \theta_6, \;\; \mu'_{4,6} = \theta_3,$$

while all other elements are equal to zero. Define also the $6 \times 6 \times 4$ array $\boldsymbol{\mu}''$ such that

$$\mu''_{i,k,1} = \mu''_{k,i,1} = \mu_1/(\theta_i \theta_k), \quad \text{for } \; i = 1, \ldots, 6, \;\; \text{and } \; k = i+1, \ldots, 6,$$

$$\mu''_{2,4,2} = \mu''_{4,2,2} = 1, \;\; \mu''_{5,5,3} = 2\theta_1/\theta_5^3, \;\; \mu''_{1,5,3} = \mu''_{5,1,3} = -1/\theta_5^2, \;\; \mu''_{3,6,4} = \mu''_{6,3,4} = 1,$$

while all other entries are equal to zero. Then, the i, k element of the Hessian of the log-posterior is

$$\{\nabla^2 \log \pi(\boldsymbol{\theta})\}_{i,k} = \{\nabla^2 \log \pi(\boldsymbol{\theta})\}_{k,i} = l''_1 \mu'_{1,i} \mu'_{1,k} + l'_1 \mu''_{i,k,1} + l''_{j(i)} \mu'_{j(i),i} \mu'_{j(i),k} + l'_{j(i)} \mu''_{k,i,j(i)} + \mathbb{1}_i(j) p''_i,$$

where $\mathbb{1}_i(j)$ is equal to 1 if $i = j$ and 0 otherwise. Notice that, while the above expression for the Hessian is very compact, this is not an efficient way of computing it.